Introduction aux sondages à l'usage du plus grand nombre

Benoît Riandey, Blöss-Widmer Isabelle

Introduction aux sondages à l'usage du plus grand nombre

Benoît Riandey (INED), Isabelle Widmer (Université de Provence)

Introduction

Nous avons eu l'occasion d'enseigner les méthodes de sondage à des publics variés allant de statisticiens démographes aux personnels administratifs de l'INED, en passant par les étudiants en sociologie ou en journalisme. Les objectifs étaient variés : faire comprendre le principe du sondage pour pouvoir en parler sans dire trop de bêtises ou acquérir un savoir-faire pour permettre une pratique professionnelle individuelle ou collective de conception d'une enquête par sondage. Les moyens à mettre en oeuvre et les pédagogies sont adaptées aux objectifs et aux publics : ceux qui ont une sensibilité de matheux veulent comprendre l'origine des formules ; les autres préfèrent qu'on ne les montre pas trop. Et si une formule bien expliquée parle d'elle-même, les fichiers pédagogiques et les simulations sont les outils de base de l'enseignement. D'ailleurs, pour un praticien expérimenté, une formule est bien souvent liée à l'histoire d'une enquête et à un choix opéré, à une erreur ou une catastrophe statistique. Rien de tel que les erreurs passées pour comprendre les choses. Certaines notions théoriques seront, en effet, comprises en faisant appel à l'imaginaire des étudiants et donc à des situations très concrètes. Or, dans l'observation de la vie quotidienne, chacun fait du sondage sans s'en apercevoir, avec les risques d'erreur d'observation liés essentiellement aux erreurs de champ et à une mauvaise perception de l'unité de sondage et donc des conditions d'inégales probabilités d'inclusion. Des paradoxes résultent de ces confusions. La réflexion en forme de devinette sur ces paradoxes est dans tous les cas une bonne introduction aux sondages.

La précision d'un sondage repose sur la notion d'intervalle de confiance bien abstraite pour certains auditoires. Un intervalle de confiance au seuil de 10 %, 5 %, 1 % n'a-t-il pas la propriété d'un coup de filet à papillons sur la vrai valeur recherchée avec 10, 5 ou 1 chances sur cent de la manquer en se servant de filets moins larges ou plus larges, de filets une étoile, deux étoiles ou trois étoiles ? De là à parler du Michelin de la statistique, il n'y a qu'un pas à franchir, ce que nous avons fait.

Le sondage est aujourd'hui une des introductions de la statistique et des probabilités particulièrement proche du public et des jeunes, plus sans doute encore que la démographie. Il a pris la relève du rôle historique des jeux de hasard. A ce titre, le sondage tient une place irremplaçable dans la vulgarisation de la statistique

1- Des idées sommaires

Nous avons tous entendu les propos les plus sommaires sur la qualité des enquêtes :

- « Un million de réponses volontaires valent mieux qu'un sondage ».
- « La méthode des quotas est meilleure car au moins les variables de quotas sont mesurées sans aucune erreur d'échantillonnage ».

- « Mon échantillon est bon : mes quotas sont parfaitement respectés ».
- Sur 40 patients, j'en ai enquêté 20 : un taux de sondage de 50 %, c'est bien ».
- « Comment se fait-il que les statistiques prétendent que les autobus sont seulement à moitié pleins alors que vous et moi montons toujours dans des bus bondés » ?
- « Les sondages politiques sont bons à 3 % près».
- « Pas représentatif mon échantillon ? pas grave, on redresse ».

Ces propos dévoilent le défi pédagogique à relever alors que communément faire un sondage, c'est demander de répondre à un questionnaire d'opinion.

2- Comprendre le hasard mathématique : notion, bienfaits et limites

La première idée à faire passer est très abstraite : hors de la collecte exhaustive, du recensement donc, seul le tirage au hasard mathématique nous protège du biais de sélection (biais sélectif). La bonne volonté ne remplace pas le hasard et le hasard est très difficile à simuler.

Qui parmi nous n'a dû jouer à écrire une suite fictive de piles ou faces avant qu'on ne lui fasse compter le nombre de séries longues de piles (ou de faces) et constater qu'il n'a pas osé en écrire un nombre suffisant. Nous sommes de bien faibles simulateurs du hasard. De grâce, laissons le faire ![1]

Au printemps 1994, le premier ministre Edouard Balladur organise une vaste consultation des jeunes âgés de 15 à 25 ans et recueille un million cinq cent trente neuf mille questionnaires. Suite à ce succès quantitatif, en meilleur politique que statisticien, il déclare que « un million cinq cent mille réponses volontaires valent mieux qu'une enquête par sondage ». Au contraire, les scientifiques nient toute valeur scientifique à ce vrac de questionnaires et la Société Française de Statistique (SFdS) demande que ce questionnaire soit soumis à un échantillon représentatif, pour disposer d'une information scientifiquement fondée. Ainsi la consultation estimait à 33 % la proportion de jeunes « *pas du tout confiants dans l'avenir* », tandis que quelques mois plus tard SCP Communication estimait cette proportion à seulement 13 %. Le décalage temporel ne rend certes pas directement comparables ces proportions, mais il conforte l'idée que l'inquiétude pour l'avenir aurait été une très forte motivation pour répondre à la consultation organisée par le Premier Ministre. Le succès de cette consultation postale était l'expression du mécontentement politique des jeunes. Ce n'est guère un critère de représentativité. Le nombre, sans méthodologie, ne suffit ainsi aucunement à faire la qualité. Le hasard mathématique est nécessaire pour protéger des biais de sélection.

Plus généralement, les statisticiens ont repéré que le biais sélectif des réponses spontanées, notamment des enquêtes par voie de presse, est souvent motivé par une réaction relevant de l'esprit de contradiction ; ce moteur psychologique s'avère facilement activé mais nous en conviendrons, il est, lui aussi, très peu favorable à la représentativité.

De nombreuses situations induisent des biais de sélection : « comme par hasard », les femmes décrochent plus souvent le téléphone lors d'enquêtes téléphoniques. En se rendant au domicile

[1] On pourra consulter a ce propos http://www.statistix.fr/spip.php?article50

des enquêtés, les enquêteurs préfèrent interroger une personne présente, donc probablement souvent à la maison, plutôt que revenir interroger une personne absente lors du premier passage. Bref, on a toujours de bonnes raisons de choisir et les autres font « comme par hasard » le plus souvent les mêmes choix. Le hasard des opportunités à saisir n'est que source de biais … mais aussi de réduction des coûts. Cruel et permanent dilemme !

Choisir ses enquêtés peut permettre dans certains cas « d'équilibrer l'échantillon ». Mais tout dépend des critères de choix. A témoin, cet exemple caricatural d'une première rencontre avec un enquêteur recruté pour des enquêtes sur quotas : « *Pour vos enquêtes, je vous sélectionne des enquêtés sérieux, pas ces gens qui racontent n'importe quoi. D'ailleurs j'ai remarqué qu'en discutant du thème du questionnaire avant l'interview avec l'enquêté, il répond beaucoup mieux au questionnaire. »* J'ai compris une nouvelle fois ce jour là l'importance de la formation des enquêteurs à la notion du hasard mathématique, comme à celle de la neutralité de l'enquêteur.

3- Univers et unité statistiques : bizarreries et paradoxes

Le tirage de l'échantillon suppose d'abord l'existence d'une base de sondage dont on scrutera l'exhaustivité, la mise à jour, la richesse informative et la proportion de la sous-population concernée dans cette base. Ces questions pratiques importantes sont directement transposées de la définition démographique d'une population avec ses trois critères d'inclusion : le concept définissant la population, ses délimitations temporelle et géographique.

Première nécessité, sensibiliser les étudiants à l'identification de l'unité de sondage, parfois distincte de l'unité statistique d'analyse. Le sondage à plusieurs degrés familiarise les étudiants avec la question, mais certainement insuffisamment tant on voit d'erreurs de professionnels dans l'usage des pondérations.

Le plus souvent, la question de la définition de l'unité statistique à enquêter ne se pose pas : un sondage électoral d'électeurs traite d'individus ; une enquête budget de famille fait appel à un sondage de ménages. Mais cette simplicité habituelle peut être trompeuse.

Heureusement, quelques paradoxes viennent attirer notre attention *: « Dans l'après-guerre un quart des mères avaient quatre enfants ou plus. Pourtant dans ma classe, comme dans les autres, nous étions la moitié à appartenir à une famille d'au moins quatre enfants* ».
Avec pédagogie, Laurent Toulemon démonte l'erreur tendant à confondre un sondage de mères pour l'étude des descendances avec un sondage d'enfants[2] pour la mesure des fratries. Certes un sondage d'enfants permet l'estimation de la descendance de leurs mères, mais en tenant compte des probabilités inégales introduites au niveau des mères précisément par ces descendances inégales ; l'erreur est conséquente. Ne confondons pas unité de sondage et unité d'analyse[3].

Le paradoxe n'était pas différent pour les statistiques (un peu anciennes) de remplissage des avions : « *Comment les compagnies aériennes s'accordent-elles à déclarer un taux de remplissage moyen de 50 % alors que quand vous ou moi prenons l'avion, il est presque*

[2] 3,1 enfants par mère de la génération 1930, mais des fratries moyennes de 4,5 pour les enfants de ces mères.
Cf. *Population et Sociétés* n° 374 (www.ined.fr/publications). Voir l'annexe pédagogique jointe.
Un fichier pédagogique de cette application est disponible à l'adresse www-ehf.ined.fr

[3] Voir aussi http://www.statistix.fr/spip.php?article25

toujours plein » ? Dans ce sondage de passagers, dont nous sommes les observateurs, chaque vol est observé avec une probabilité proportionnelle à son nombre de passagers. Pondérez donc vos observations par l'inverse de ce nombre de passagers et vous tomberez d'accord avec les compagnies.

Cet exemple illustre comme notre vision du monde ambiant peut être biaisée par l'ignorance de nos probabilités d'observation et nous oriente vers une nouvelle disposition mentale pour reconstituer cet univers ; par exemple, désormais, quand vous roulerez à 125km/h sur l'autoroute et que vous verrez dépassé par une multitude de Mercedes et de BMW, vous ne penserez plus que le parc automobile français est essentiellement constitué de ces voitures, mais, pondérant chaque voiture qui vous dépasse (ou que vous dépassez) par l'inverse de la différence de vitesse des deux véhicules, vous aurez une vision juste du parc automobile empruntant l'autoroute ! (Cependant vous n'observerez pas les voitures roulant comme vous à 125 km/h, mais devrez changer d'allure pour les observer et corriger ainsi cette erreur de couverture).

Vous aurez pensé comprendre pourquoi telle enquête sur la fréquentation des autobus indique des autobus bondés alors que là encore les entreprises de transports annoncent un taux de remplissage de 50 %. Et bien non, il ne s'agissait pourtant pas d'un sondage de passagers, mais bien d'autobus : il avait été décidé de tirer le prochain autobus se présentant à un arrêt intermédiaire donné. Pourquoi donc malgré cette précaution de sondage, rencontrer la même estimation biaisée ? Nous présentant au hasard un instant donné, nous tirons un autobus avec une probabilité proportionnelle à l'intervalle inter-bus. Nous tirons donc avec une probabilité plus grande les bus qui se font attendre et que nous savons bondés. Bref nous avons tiré des secondes d'attente et non directement des bus, d'où l'estimation biaisée. Tirons donc nos bus dans la liste de la compagnie, ou pondérons par l'inverse du temps d'attente… en laissant passer le premier bus pour observer l'intervalle d'attente complet[4].

Cette anecdote nous rend apte à l'exercice pédagogique suivant, histoire vraie ou inventée dont on ne connaît pas l'auteur. Avant le développement de la gestion informatique, une société de logements sociaux s'inquiétait de la fréquence apparente des impayés. Le responsable recrute un stagiaire et lui dit : « Tu vas me tirer cent dossiers au hasard et me les regarder de près ». Le stagiaire, observant 75 mètres linéaires de dossiers classés par ordre alphabétique, mesure 75 centimètres sur une ficelle (ou avec un mètre de couturière) et sort un dossier par longueur de ficelle.

« Une telle proportion de dossiers d'impayés, ce n'est pas possible ! » lui rétorque son chef à la présentation des résultats. Où est l'erreur ?
Le stagiaire n'a pas réalisé un tirage uniforme entre dossiers, mais a tiré des centimètres et donc des dossiers, proportionnellement à leur épaisseur ; bien sûr les dossiers avec impayés sont les plus épais.
Qu'aurait-il fallu faire ?
Puisque les dossiers sont classés par ordre alphabétique, non signifiant, les dossiers successifs sont d'épaisseur indépendante. Ayant mesuré 75 centimètres, il aurait suffi de prendre le dossier suivant.
Que faut-il faire maintenant avec ces 100 dossiers sortis sur la table ?
Le stagiaire reprend son mètre de couturière, mesure l'épaisseur de chaque dossier, le pondère par l'inverse de son épaisseur.
Bravo 20 sur 20 !

[4] Nicolas Gauvrit « Statistique, méfiez- vous » (2007), p.192, Ellipse.

Comment 20 sur 20 ? Ce n'était aucunement un sondage au hasard, mais un sondage systématique dont la première unité tirée détermine toutes les autres par simple progression arithmétique !

Vous avez raison, mais le **sondage systématique**, ce n'est pas si mal ; c'est facile et il offre de multiples usages. D'ailleurs nous faisons régulièrement des tirages systématiques sans le vouloir, par exemple en distribuant les cartes, au bridge notamment[5]. Dès avant distribution, les quatre jeux de 13 cartes sont constitués mais pas attribués. En coupant le jeu (soit en tirant au hasard un nombre, demeuré inconnu, entre 1 et 4 ou entre 0 et 3), on vous affecte votre échantillon de treize cartes au taux de sondage ¼ selon la règle de la progression arithmétique.

Si le jeu a été bien battu, on se convainc facilement de ce qu'un tirage systématique est alors équivalent à un sondage aléatoire simple.

Si le jeu est trié selon une catégorie qualitative ou quantitative, le tirage systématique équivaut à une quasi stratification selon ce critère[6]. C'est précisément ce qui se passe avec un jeu neuf dont les cartes sont triées par valeur décroissante de l'as au 2, les quatre cartes de la strate respectant l'ordre pique, cœur, carreau, trèfle. Quelle que soit la coupe, vous recueillerez un jeu parfaitement équilibré entre les treize strates (valant toujours 10 points d'honneur au bridge). Sur un fichier opportunément trié, le sondage systématique est meilleur que le sondage aléatoire simple.

De plus, vous apprendrez là que le tirage systématique est sujet à une régularité bizarre quand le pas de sondage est multiple d'un cycle présent dans la base. C'est le cas qui nous occupe puisque avec le jeu neuf, un joueur recueillera treize cartes de la même couleur et serait ainsi porté à dire « *D'après mon échantillon, ce jeu ne comporte que des piques* ». Cette bizarrerie se retrouve lorsqu'à un mariage, dans une farandole bien réglée alternant garçons et filles, vous tirez un échantillon de danseurs avec un pas de sondage pair : à la seule vue de l'échantillon vous jugeriez la noce entièrement masculine ou féminine. Toutefois, cette situation a peu de chance de se manifester dans la pratique professionnelle où le pas de la progression est généralement décimal.

L'exemple des cartes ne saurait illustrer une performance majeure du sondage systématique : cette technique permet, en effet, de tirer des communes (des unités primaires) proportionnellement à leur population. Il suffit de cumuler les populations des communes - opportunément classées par population croissante- et de sélectionner la commune d'un individu fictif tous les n individus. Ce glissement entre unités statistiques peut prendre à défaut le sondeur s'il n'y a pris garde, comme l'illustre la mésaventure réelle suivante.

La **méthode des itinéraires** désigne une application classique du sondage systématique. Il s'agit d'opérer un sondage de logements en l'absence de fichier. Le tirage est alors opéré directement sur le terrain : on définit un balayage de l'espace sondé, par exemple le quartier - l'Iris dirait l'INSEE - qui ne passe qu'une fois à chaque adresse ; puis l'enquêteur retient un logement tous les n.

Cette méthode est très employée aux Etats-Unis dans un habitat pavillonnaire. En France,

[5] Au tarot, la distribution des jeux n'est pas systématique du fait de la constitution du chien.

[6] Le « quasi » tenant à des taux de sondage décimaux, ce qui n'est pas le cas de l'exemple.

elle a donné lieu à quelques belles catastrophes statistiques. Ainsi tel étudiant géo-démographe, enquêtant en habitat mixte, sélectionnait une adresse sur 7 et, à sa propre surprise, conclut de son enquête que la commune populaire enquêtée était bien loin d'atteindre le seuil des 20 % de logements sociaux requis par la loi SRU relative au logement social. Grossière erreur méthodologique, mais laquelle ?

Ce tirage systématique a été mené dans un univers d'adresses et non de logements. Ainsi un pavillon avait une chance sur 7 d'être tiré ; un logement social d'un HLM de 30 logements avait une chance sur 210 de figurer dans l'échantillon.
Que fallait-il faire ? *Cumuler le nombre de logements à chaque adresse et, au cours de l'itinéraire, tirer des logements et non des adresses. Ainsi, dans ce HLM, on aurait, selon le hasard, sélectionné 4 ou 5 logements*.
Que devait faire l'étudiant une fois son erreur repérée ? *Pondérer chaque observation par la variable* `Nombre de logements dans le même immeuble`*, puis mesurer correctement la variance de l'estimation*.
Parfait, la leçon est bien retenue, mais la dimension des immeubles pavillonnaires et sociaux est tellement disproportionnée que la variance explose.
Que faut-il alors faire *? Retourner dans le quartier et faire un sondage complémentaire dans les immeubles collectifs, puis fusionner correctement les deux sous-échantillons*.

En fait, il aurait fallu cumuler le nombre de logements à chaque adresse en parcourant le quartier et retenir un logement tous les n, ce qui n'est pas du tout commode à faire, même en connaissant a priori l'effectif des logements du quartier. Dans la pratique, les communes de plus de 10 000 habitants tiennent à jour un répertoire des immeubles localisés, le RIL du *Recensement rénové*. Ce répertoire enregistre le nombre de logements à chaque adresse. Il n'y aurait aucune difficulté technique à procéder à un sondage de logements dans ce fichier anonyme. Pourquoi y fait-on obstacle ? De même, le bulletin immeuble du recensement des petites communes dénombre les logements de l'adresse. En habitat aggloméré, il ne devrait pas y avoir besoin de recourir à la méthode des itinéraires en France.

4- Précision d'un sondage

L'idée (fausse) la plus répandue sur la précision des sondages est qu'elle est fonction du taux de sondage. La précision d'un sondage ne dépend pas du taux de sondage, mais de la taille de l'échantillon[7]. Ainsi la formule approchée de la précision d'une proportion parle : l'expression $\pm 2\sqrt{pq/n}$ dit bien que le taux de sondage n'influe pas. Est chassée l'opinion rassurante « Un taux de sondage de 50 % ou de 10 %… c'est bien ».

Vient le moment de faire jouer nos auditeurs avec le joli tableau de Anne-Marie Dussaix et Jean-Marie Grosbras[8] reproduit à la suite : « *Pour avoir une enquête deux fois plus précise, il faut un échantillon quatre fois plus grand. Pour gagner une décimale de précision, il faut un échantillon cent fois plus grand, cent fois plus cher. Pour estimer avec une précision de 2 % (0,02) une proportion tournant autour de 10 % (0,10), l'échantillon est de taille 900 ».*

[7] Pour une population donnée de taille définie connue, préciser le taux de sondage ou la taille de l'échantillon sont synonymes, mais cette situation tautologique est tout autre.
2 Dussaix A.M. et Grosbras J.M., « Les sondages. Principes et méthodes » (1993), p. 27 Que sais-je, n° 701.

Tableau 2.2. – **Précision de l'estimation d'une proportion calculée à partir d'un échantillon**

Proportion observée p / taille d'échantillon n	5 % ou 95 %	8 % ou 92 %	10 % ou 90 %	15 % ou 85 %	20 % ou 80 %	25 % ou 75 %	30 % ou 70 %	35 % ou 65 %	40 % ou 60 %	50 %
100					8	8,6	9,2	9,6	9,8	10
150				5,7	6,4	6,9	7,3	7,6	7,8	8
200			4,3	5,1	5,7	6,1	6,5	6,8	6,9	7,1
250	2,8	3,4	3,8	4,5	5	5,4	5,8	6	6,2	6,3
300	2,5	3,1	3,5	4,2	4,6	5	5,3	5,6	5,7	5,8
350	2,3	2,9	3,2	3,8	4,2	4,6	4,9	5,1	5,2	5,3
400	2,2	2,7	3	3,6	4	4,3	4,6	4,8	4,9	5
500	2	2,4	2,7	3,2	3,6	3,9	4,1	4,3	4,4	5
600	1,8	2,2	2,4	3	3,3	3,5	3,8	3,9	4	4,1
700	1,7	2,1	2,3	2,7	3	3,3	3,5	3,5	3,7	3,8
800	1,5	1,9	2,1	2,5	2,8	3	3,2	3,3	3,4	3,5
900	1,5	1,8	2	2,4	2,7	2,9	3	3,1	3,2	3,3
1 000	1,4	1,7	1,8	2,3	2,5	2,7	2,9	3	3	3,1
1 500	1,2	1,4	1,5	1,9	2,1	2,3	2,4	2,5	2,6	2,6
2 000	1	1,2	1,3	1,6	1,8	2	2,1	2,2	2,2	2,3
3 000	0,8	1	1,1	1,3	1,4	1,5	1,6	1,7	1,8	1,8
5 000	0,6	0,8	0,8	1	1,1	1,2	1,3	1,4	1,4	1,4
10 000	0,4	0,5	0,6	0,7	0,8	0,9	0,9	1	1	1

La simple consultation de ce tableau nous débarrassera déjà de simplifications médiatiques statuant sur la précision des sondages politiques du type : « Avec 20 % d'intentions de vote, on connaît le pronostic à 3 % près pour ce candidat ». En fait, avec un échantillon de 1000 répondants, la largeur de l'intervalle de confiance n'est pas toujours de ± 3 %, mais au maximum de ± 3 %, et à ce niveau de fréquence de ± 2,5 %. Voilà qui serait compliqué à mentionner dans la grande presse.

Nos collègues objecteront à juste titre que cette présentation de l'intervalle de confiance n'est pas juste **:** la formule présentée relève du sondage aléatoire simple avec remise, mais cette notion de remise n'a pas de sens pratique dans les sondages puisqu'on ne va pas poser deux

fois le même questionnaire à une personne. Patience, on ne peut purger proprement deux idées fausses en une parole. Nous y reviendrons avant deux minutes.

Dans la situation classique des sondages nationaux, la formule rigoureuse des sondages sans remise $1{,}96\sqrt{1-f}\sqrt{pq/n}$ est tellement numériquement identique à celle du sondage avec remise qu'on évitera aux non mathématiciens d'entrer dans cette distinction troublante.

Cette première étape passée, on introduira dans un second temps la formule exacte comme une correction nécessaire fondée sur le contre-exemple flagrant du recensement, puis on examinera jusqu'à quel taux cette correction s'impose. On éveillera, en effet, l'esprit critique de nos auditeurs sur ce fait : par nature**,** l'incertitude d'échantillonnage des estimations d'un recensement est nulle ; or, la formule approximative $2\sqrt{pq/n}$ ne le montre pas. Il est donc temps d'introduire le facteur correctif $\sqrt{1-f}$ qui précisément s'annule pour la valeur 1 du taux de sondage f aux recensements.

Ce facteur correctif pèse lourd quand les taux de sondage sont inhabituellement très élevés (1/10ème, 1/4, 1/2). Un auditoire sceptique en prendra conscience avec l'exemple suivant : une administration en charge populations étrangères lance un appel d'offre pour des études ne devant pas dépasser 100 interviews. Un chercheur remporte l'appel d'offre avec un projet concernant un corps professionnel 130 personnes étrangères. Malgré ce taux de sondage de ¾ environ (77%), l'incertitude d'échantillonnage n'est réduite que de moitié par rapport à un échantillon de même taille en population générale[9]. Conclusion du statisticien ? *« Une bouteille à moitié pleine est encore à moitié vide. Donc vous êtes en bonne voie, terminez le travail en enquêtant les 30 derniers individus* ». Dans les petites populations, le seul bon sondage est l'enquête exhaustive, le recensement. Au jeune médecin satisfait d'avoir enquêté 20 de ses 40 patients, nous répondrons toujours : « Vous êtes en bonne voie, terminez le travail en enquêtant les 20 autres ».

5- Complexe la notion d'intervalle de confiance ?

Revenons sur l'énoncé amélioré : « On connaît à 2,5 % près une intention de vote voisine de 20 % ».

Malgré tout, un lecteur rigoureux serait avisé d'objecter « *Pas du tout, dans cinq cas sur 100, la valeur à estimer ne figurera pas dans cette fourchette* ». Que comprendra le lecteur si on lui assène que la fourchette laisse passer la valeur cherchée pour 5% des échantillons ?

On lui expliquera qu'en sélectionnant un échantillon au hasard dans la population, même d'une taille importante de 1000 personnes, on peut avoir la malchance insigne de sélectionner uniquement des électeurs aficionados d'Arlette et le lecteur comprendra que les sondages sont susceptibles de se tromper, mais l'explication est insuffisante. D'où sortent ces 5 % ?

Nous vous proposons d'illustrer cette énigme par une métaphore de chasse aux papillons.

[9] $\sqrt{(1-3/4)} = \sqrt{1/4} = 1/2$

Une estimation d'enquête
=
Un coup de filet à papillons

avec - 1 chance sur 10 de rater la cible (estimation *)
- 1 chance sur 20 de rater la cible (estimation **)
-1 chance sur 100 de rater la cible (estimation ***)

La valeur à estimer est notre papillon. Un filet à papillons est doté d'un point central et d'une largeur autour de ce centre (d'un rayon s'il est circulaire). De plus, le filet à papillons rate d'autant plus souvent sa cible qu'il est plus étroit. Le filet à papillons est donc doté de toutes les caractéristiques définissant un intervalle de confiance. Le statisticien qui a collecté son enquête dispose d'un jeu de filets à papillons d'une, deux ou trois étoiles, de largeurs différentes lues sous la courbe de Gauss. Ils manquent respectivement le papillon 1 fois sur 10, 1 fois sur 20 ou 1 fois sur 100. Le filet trois étoiles, beaucoup plus sûr avec seulement un échec sur cent, est beaucoup plus large et attrapera beaucoup plus de papillons qu'un filet moins large et parmi les papillons attrapés, on ne sait pas quel est le bon. En fait, plus le filet est grand, plus on a de chances d'attraper le bon papillon, mais plus il y en aura dans le filet. Le filet trois étoiles est donc à la fois assuré de succès, mais de résultat imprécis.

Une estimation d'enquête est donc un coup de filet à papillons qui prendra sa proie… sauf x fois sur cent. Pour ne pas se compliquer la vie, les statisticiens prennent, sauf mention contraire, un filet deux étoiles. Ils s'arrogent donc le droit à l'erreur 1 fois sur 20.

Quoiqu'il en soit de la théorie, sa mise en œuvre en France est plutôt limitée à la statistique publique et à la recherche. Mais elle sert de cadre de référence et d'évaluation aux enquêtes empiriques des instituts privés, notamment aux enquêtes sur quotas qu'il convient maintenant d'examiner.

6- Derrière les quotas.

Le mot quota est connu de longue date dans les médias et a d'abord renvoyé aux fameux « quotas laitiers ». De fait, il s'agit simplement d'un effectif de référence et aucune nuance ne permet d'y associer un principe de sélection des unités constituant cet effectif. A ce titre, dans les sondages, ce mot est en lui-même une source de malentendus comme le mot hasard, ou le mot sondage employé pour enquête d'opinion par *sondage*. Un travail sur les définitions et le vocabulaire s'impose à nouveau avant tout contenu technique.

« *Mes quotas sont vérifiés, j'ai un bon échantillon* ».

Il est exact que des quotas corrects améliorent l'estimation des variables très corrélées aux variables sous quotas, par exemple le risque de chômage en fonction de la pyramide des âges. Mais d'abord les variables des quotas sont-elles aussi bonnes qu'elles l'apparaissent ? Souvent les ouvriers manquent dans l'échantillon, mais dans le cas contraire, ne recueille-t-on pas une proportion excessive d'ouvriers qualifiés ? Que dire des cadres supérieurs ? Ne sont-ils pas en trop grande proportion des personnels du secteur public, notamment des professeurs ? A supposer que les variables des quotas soient intrinsèquement satisfaisantes,

elles n'apportent de garantie que pour elles mêmes. Que dire d'un échantillon de mères fondé sur des quotas strictement respectés dans une enquête sur le thème de la garde des enfants, mais collectée par facilité à la sortie des crèches ? « *100 % des nouveaux-nés sont gardés en crèches prouve cette belle enquête sur quotas* » ! Les quotas nickel n'apportent aucune garantie à l'encontre des biais d'échantillonnage (probabilité forte des mères des bébés gardés en crèche et nulle ou négligeable pour les autres mères).

7- Pourquoi les enquêtes par quotas, ça marche souvent ?

Il ne faut pas se faire d'illusion : les individus n'ont aucunement des probabilités égales d'être enquêtés par quotas. La rigueur voudrait que leur observation soit pondérée par l'inverse de leur probabilité d'inclusion … bien inconnue. Mais le biais constitutif des échantillons sur quotas n'entraîne pas nécessairement des estimations biaisées : le théorème de Horwitz-Thomson indique qu'il suffit que les probabilités d'inclusion -inconnues- soient indépendantes de la variable mesurée pour que l'estimation soit non biaisée. Dans ce cas, le biais sélectif n'est pas perturbant. Mais cette hypothèse n'est vérifiable que sur des variables déjà mesurées par ailleurs. Sinon l'expérience du terrain et le bon sens suppléent plus ou moins bien au contrôle de l'hypothèse. La statistique est complaisante, cette hypothèse se vérifie assez souvent pas trop mal.

Cette réflexion théorique sur les estimations correctes à partir d'échantillons biaisés s'adresse à des spécialistes. Après avoir convaincu les étudiants qu'il faut solliciter l'action du hasard, leur laisser entendre qu'on peut très bien réaliser de bonnes enquêtes en ne tenant aucun compte de ce principe fondamental devrait dérouter les étudiants autant que leur malheureux enseignant. Enseigner une chose et son contraire n'est pas confortable. L'argument selon lequel « avec de petits échantillons, le sondage aléatoire n'est pas très bon » vient à la rescousse de l'enseignant qui complètera en ajoutant que « rien ne garantit que la qualité des échantillons par quotas s'améliore en accroissant son échantillon exposé au biais».

8- Pourquoi les quotas ne marchent pas toujours ?

Les biais de sélection sont souvent liés à la variable mesurée. Aussi est-on attentif à éviter les refus thématiques. Ainsi, lors d'une enquête par quotas sur la pratique sportive, les enquêteurs recevaient une première consigne absolue : ne pas prononcer le mot « sport » avant l'acceptation de l'enquêté. On visait par là à éviter des désistements sélectifs: *« Posez votre questionnaire à mon frère ; il répondra beaucoup mieux que moi.* » Le taux de participation sportive inattendu à cette enquête tient-il au non respect de cette consigne, au recrutement des enquêtés à la sortie de salles sportives, à un recrutement d'enquêteurs sportifs, à une définition large non contrôlée de la pratique sportive ? Personne ne pourra le dire.

Aussi, quelle meilleure introduction à une enquête que celle peu inductive des sondages omnibus « *Voulez-vous bien répondre à un questionnaire sur des thèmes variés* » ?

9- Les quotas plus précis que le sondage aléatoire ?

Oui, ça se peut. Des quotas bien choisis réduisent la variance de l'estimation, l'amplitude de son intervalle de confiance. Si l'échantillon est sans biais l'échantillon sur quotas sera plutôt plus précis qu'un sondage aléatoire simple. Reste à espérer qu'une main invisible ait bien guidé les enquêteurs. L'idéal serait d'avoir un échantillon aléatoire avec quotas. C'est le

sondage équilibré. Il bénéficie des avantages de l'aléatoire et des quotas, mais bien sûr ce ne sont pas les enquêteurs qui désignent les enquêtés. Il s'agit d'équilibrer l'échantillon sans le libre choix des enquêtés.

10- Redressement

« Pas représentatif mon échantillon ? Pas d'importance, on redresse ». Cette réflexion ancienne d'une responsable d'un grand institut de sondage français n'avait pas manqué de choquer son interlocuteur, mais elle pose une question théorique et pédagogique importante et complexe. Elle comprend sa part de vérité comme d'erreur.

C'est vrai qu'il est facile de remettre à niveau la proportion d'hommes ou de Franciliens dans un échantillon. Mais est-il indifférent que l'algorithme de redressement, expression déplorable, corrige par post-stratification une fluctuation aléatoire ou masque un biais issu d'un phénomène de sélection ? La réponse dépend du processus sélectif. Si spontanément les femmes répondent davantage aux enquêtes, c'est pour une part parce qu'elles sont plus souvent présentes au logement. Cette cause signifie que les femmes au foyer seront sur-représentées par rapport aux femmes actives. Et que ce soit par redressement ou par quotas que le sous-échantillon féminin est remis au bon niveau ne change rien à son biais de structure.

Ainsi, même redressés de leur sous-représentation des catégories ouvrières, les panels postaux des instituts privés conservent le biais intrinsèque du mode de collecte à l'encontre des personnes peu instruites, mal à l'aise avec l'écriture. « *Il n'y a qu'à redresser en fonction du diplôme* » ! Hélas, le diplôme est la variable la plus mal renseignée au recensement, même pour les Français de naissance. L'importance de ce biais ne peut qu'être approchée faute de statistique fiable, si ce n'est au niveau national par l'enquête Formation Qualification Professionnelle de l'INSEE[10].

Un excellent taux de réponse est sensé garantir l'absence de processus sélectif et l'INED s'abstenait de redresser les échantillons de ses enquêtes collectées par l'INSEE avec un taux de succès de 90 %. Néanmoins, l'ancienne enquête Emploi, du haut de ses 94 % de succès, enregistrait cependant une sous-représentation de certaines catégories de jeunes peu présents au domicile. La meilleure des enquêtes en France demeurait donc imparfaite.

Revenons au mot **redressement**. Il décrivait l'enchaînement de « vilaines règles de trois » qui redonnait une allure propre aux échantillons des instituts privés … ou de l'INED, tandis que l'INSEE *repondérait* ses échantillons grâce une superbe inversion de matrice. Nous devons à Jean-Claude Deville d'avoir effacé cette représentation : l'algorithme très simple de ratissage alternatif inventé par Deming est une optimisation très légitime des pondérations sous la contrainte du respect de distributions statistiques. « *Pourquoi faire compliqué quand on peut faire aussi bien simplement* ? » nous taquinait-il avant qu'Olivier Sautory conçoive son programme de redressement Calmar. Il reste que l'appréciation d'un sondage atteint d'une forte proportion de non réponses suscite un malaise du professionnel : le redressement ne fournissant une estimation solide que si les hypothèses d'indépendance locale à la base du

[10] Claude Thélot (2004), « Tel père, tel fils ? », nouvelle édition, Pluriel.

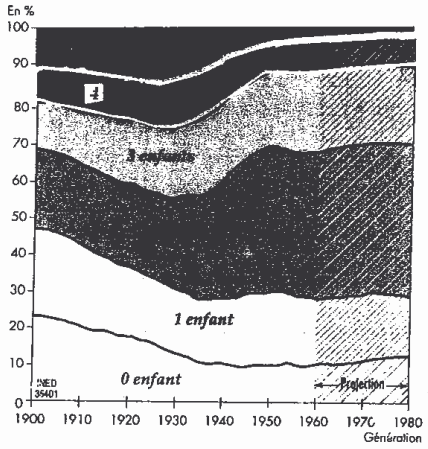
redressement ne sont pas trop fausses. Les redressements ne corrigent pas tous les maux d'un trop mauvais échantillon.

Conclusion

En conclusion, on peut comprendre les notions de base de la théorie des sondages sans entrer dans les développements techniques et les calculs spécifiques**.** Par ailleurs, les logiciels produisent des traitements d'enquêtes faites quasiment "en pilotage automatique", mais la pratique presse bouton qui va avec comporte des risques….

Laurent Toulemon

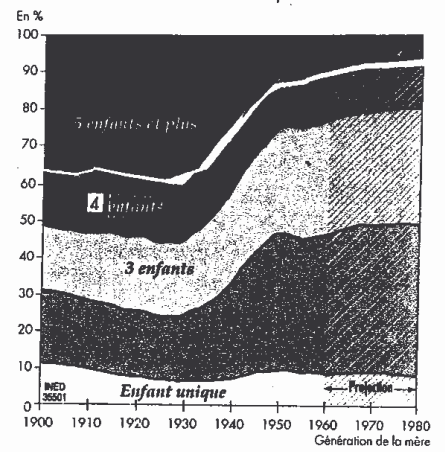
Figure 3 – Répartition des femmes selon le nombre final d'enfants



Note : La descendance par rang n'est connue de manière certaine que pour les femmes de plus de 50 ans, et avec une imprécision faible pour les femmes de plus de 40 ans (nées avant 1960). Pour les générations suivantes, la projection est fondée sur les taux de fécondité par rang de naissance des années 1996-1998 et suppose que la fécondité va rester stable à l'avenir.

Sources : Insee, état civil [2] [4], enquêtes Famille 1982, 1990 et 1999, recensement de 1946 et projections [5] [6] [9].

Figure 4 – Répartition des enfants selon le nombre d'enfants que compte leur fratrie (eux-mêmes compris)



Sources : voir figure 3.

Tableau 1 – Répartition de 100 femmes et de 100 enfants selon la taille de la fratrie, pour trois générations de femmes

Nombre d'enfants	Femmes nées en...			Taille de la fratrie (1)	Enfants dont les mères sont nées en...		
	1900	1930	1960		1900	1930	1960
0	23	13	10	0	0	0	0
1	24	18	18	1	11	7	9
2	22	26	40	2	21	19	38
3	13	18	22	3	18	20	32
4	7	10	7	4	14	16	12
5	4	6	2	5	10	11	5
6 et +	7	9	1	6 et +	26	27	4
Total	**100**	**100**	**100**	**Total**	**100**	**100**	**100**
Moyenne	*2,1*	*2,7*	*2,1*	*Moyenne*	*4,2*	*4,5*	*2,8*

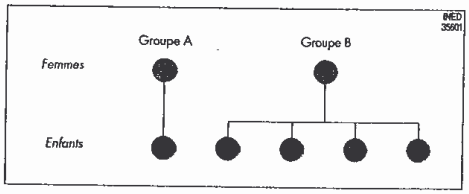
(1) Nombre de frères et sœurs y compris l'individu de référence.

Sources : voir figure 3.

Encadré 3

La taille de la fratrie : la famille vue par les enfants

Prenons 100 femmes A ayant chacune 1 enfant et 100 femmes B ayant 4 enfants. Les deux groupes de femmes sont numériquement à égalité, alors que l'on compte 4 fois plus d'enfants dans les fratries B que dans les fratries A (400 contre 100). De façon générale, le poids des familles nombreuses est plus élevé parmi les enfants qu'il ne l'est parmi les femmes. C'est pourquoi la taille de la fratrie (le nombre moyen de frères et sœurs y compris l'individu de référence) dans une population est supérieure au nombre moyen d'enfants chez les femmes ayant donné naissance à cette population.



	Nombre d'enfants			Nombre moyen d'enfants ou taille moyenne de la fratrie
	1 (groupe A)	4 (groupe B)	Ensemble	
Femmes	100	100	200	2,5
Enfants	100	400	500	3,4